Regression of Auckland House Prices

Vandol Ton, July 2020

Executive Summary

The dataset consists of property addresses along with information about the property including the number of bedrooms, bathrooms, land area and capital value and the number of people across all the age groups in the unit area based on the 2018 census.

The analysis is based on 1051 observations with 15 attributes. In the dataset, I have added two more attributes. One is the 2018 population count for each particular unit area, and the other is the 2018 index of deprivation ranging from 1 to 10, where 1 represents unit areas with the least deprived scores and 10 for the areas with the most deprived scores.

After exploring and cleaning the dataset, examining summary and descriptive statistics, creating pairs plots and visualisations of the correlation between each numerical attribute, several correlated attributes were found.

There were properties with missing number of bathrooms. I predicted the number of bathrooms based on the other attributes. I used the three most correlated attributes as they would provide the most explanatory power over the number of bathrooms in a property. The correlation of the other attributes was relatively close to zero and would not make a noticeable effect on the analysis.

After cleaning the dataset, I used Linear Regression and Random Forest to predict the capital value of a property based on the other attributes. The R^2 value of this model is not quite high and more work is required to improve the results.

Initial data analysis

The initial exploration of the data began with some summary and descriptive statistics for each of the numerical attributes.

	Bedrooms	Bathrooms	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	C18_CURPop	NZDep2018
count	1051.000000	1049.000000	1.051000e+03	1051.000000	1051.000000	1.051000e+03	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000	1051.000000
mean	3.777355	2.073403	1.387521e+06	-36.893715	174.799325	7.006319e+06	47.549001	28.963844	27.042816	24.125595	22.615604	29.360609	181.230257	5.063749
std	1.169412	0.992985	1.182939e+06	0.130100	0.119538	2.591262e+03	24.692205	21.037441	17.975408	10.942770	10.210578	21.805031	72.087700	2.913471
min	1.000000	1.000000	2.700000e+05	-37.265021	174.317078	7.001130e+06	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	3.000000	1.000000	7.800000e+05	-36.950565	174.720779	7.004416e+06	33.000000	15.000000	15.000000	18.000000	15.000000	18.000000	138.000000	2.000000
50%	4.000000	2.000000	1.080000e+06	-36.893132	174.798575	7.006325e+06	45.000000	24.000000	24.000000	24.000000	21.000000	27.000000	171.000000	5.000000
75%	4.000000	3.000000	1.600000e+06	-36.855789	174.880944	7.008384e+06	57.000000	36.000000	33.000000	30.000000	27.000000	36.000000	210.000000	8.000000
max	17.000000	8.000000	1.800000e+07	-36.177655	175.492424	7.011028e+06	201.000000	270.000000	177.000000	114.000000	90.000000	483.000000	792.000000	10.000000

There appears to be some outliers where a property consists of 17 bedrooms and a property that consists of 8 bathrooms. Furthermore, the unit areas with the maximum for the different age groups in different unit areas is a lot higher than the majority of the data. It also appears that different unit areas have different population sizes than others which have caused these outliers. I have decided to keep the extreme values in the analysis, as it appears to be real properties.

The capital value amounts are not on the same scale with the other attributes and is skewed. It makes sense to log transform the capital values to have our attributes on the same scale.

Land area was converted to an integer as it was not included in the summary and descriptive statistics table above. This involved removing the "m^2" string characters from the land area.

Lastly, I checked for any missing values as it could cause problems with the model.

	Bedrooms	Bathrooms	Address	Land area	CV	Latitude	Longitude	SA1	0-19 years	20-29 years	30-39 years	40-49 years	50-59 years	60+ years	Suburbs	C18_CURPop	NZDep2018
309	4	NaN	14 Hea Road Hobsonville, Auckland	214	1250000	-36.798371	174.647430	7002267	60	66	60	24	24	18	Hobsonville	246	2.0
311	4	NaN	16 Hea Road Hobsonville, Auckland	245	1100000	-36.798371	174.647430	7002267	60	66	60	24	24	18	Hobsonville	246	2.0
568	1	1.0	14 Te Rangitawhiri Road Great Barrier Island, ...	2141	740000	-36.197282	175.416921	7001131	27	6	6	18	39	60	NaN	156	9.0

Observations 309 and 311 have missing bathrooms, and observation 568 is missing a suburb. I have dropped observation 568 from the analysis and used linear regression to find the missing values for the number of bathrooms.

The three most correlated variables (Bedrooms, CV and NZDep2018) to bathrooms was used to build the model to predict the number of bathrooms from the properties with missing bathrooms. All the other attributes have a negligible effect on predicting the number of bathrooms.

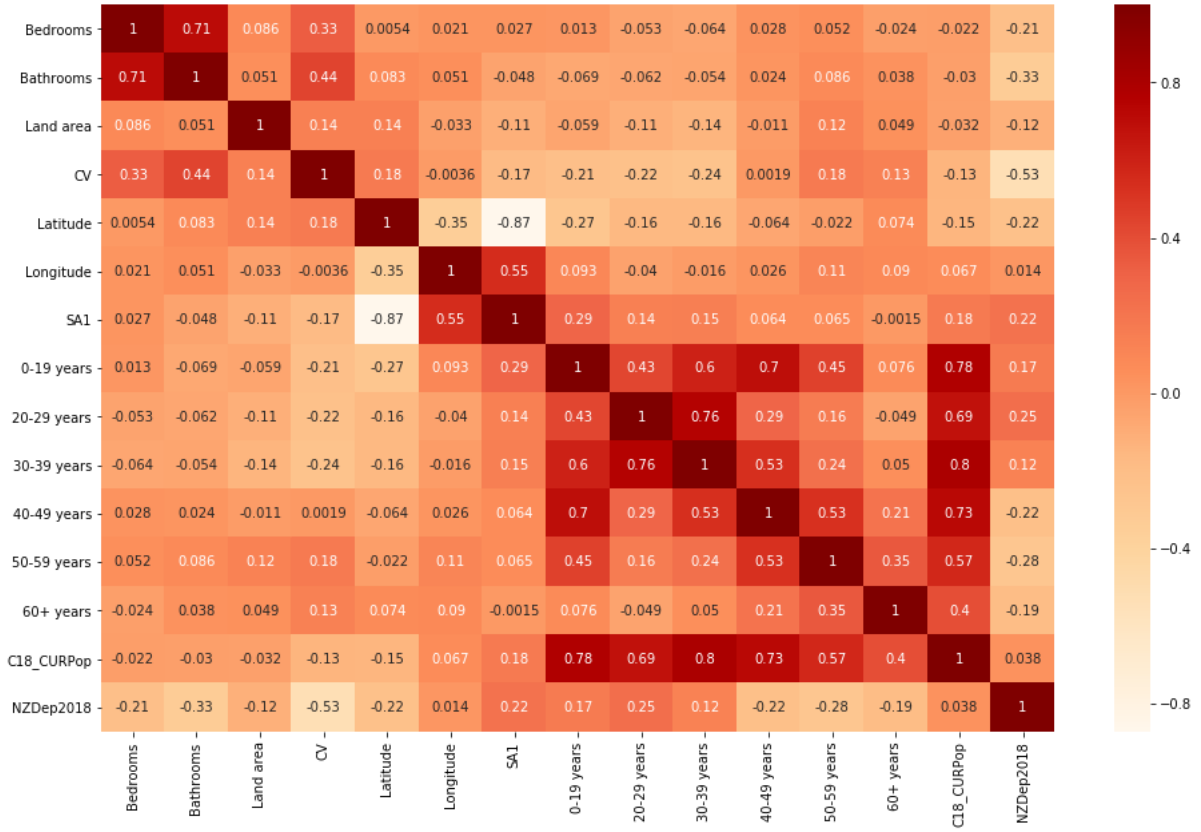
The dataset was split into a 70% training set and a 30% testing set. Linear Regression was trained on the training dataset.

The R-squared value is 0.56. Approximately over half of the observed variation can be explained by the number of bedrooms, capital value and the 2018 index of deprivation.

I used this model to predict the number of bathrooms for the ones with missing values. The Linear Regression model predicted the number of bathrooms for both observations to be 2 (rounded to the nearest whole number).

Analysis of correlations and patterns in the data

The correlation between the numeric attributes is shown below. The right bar indicates the correlation values.



The 2018 population count for the particular unit area (C18_CurPop) is correlated towards the number of people in each of the age groups. This is because the 2018 population count is the total number of people living in the unit area (excluding those who did not fill in their ages).

The 2018 index of deprivation is most correlated with the capital value, and then the number of bathrooms. All other attributes have a weak correlation with the 2018 index of deprivation.

The number of bedrooms attribute is the most correlated with the number of bathrooms. Capital value and NZDep2018 is moderately correlated with the number of bathrooms.

Model construction

The numeric attributes were kept when building the machine learning model. The dataset was split in 70% training and 30% training. Linear Regression and Random Forest were trained on the dataset.

The R-squared value for Linear Regression is around 40% while for Random Forest it is around 54%. The model using the Random Forest regressor can better explain the capital value based on the numeric attributes from this dataset.

Conclusion

Random Forest can moderately predict the capital value of properties in Auckland.

Version History

Version No.	**Changelog**
1.2	Improved flow of the report
1.1	Update to include models to predict house prices
1.0	Initial release